Homelessness Analysis Communication

Sam Farias

Intro

In our project, we will use data science methodology to study homelessness in the United States. Our motivation comes from the HUD study's goals of understanding the key factors that contribute to homelessness at the community level. HUD's objectives are: (1) to identify market factors that have a significant impact on homelessness, and (2) to create and evaluate empirical models that can accurately represent and predict homelessness in communities. We will explore if there are better modeling approaches than those described in the HUD report. Our project aims to provide valuable insights to support ongoing efforts to end and prevent homelessness in the U.S.

The overall problem is to determine whether and how housing market factors are related to homelessness. We are investigating the potential of incorporating interactions among predictor variables to develop an effective model for predicting rates of homelessness.

Our specific investigation aims to determine the effectiveness of incorporating regional differences and interactions between a geographical region variable and other predictor variables in predicting rates of homelessness. By exploring these factors, we seek to develop a model that accurately predicts homelessness rates.

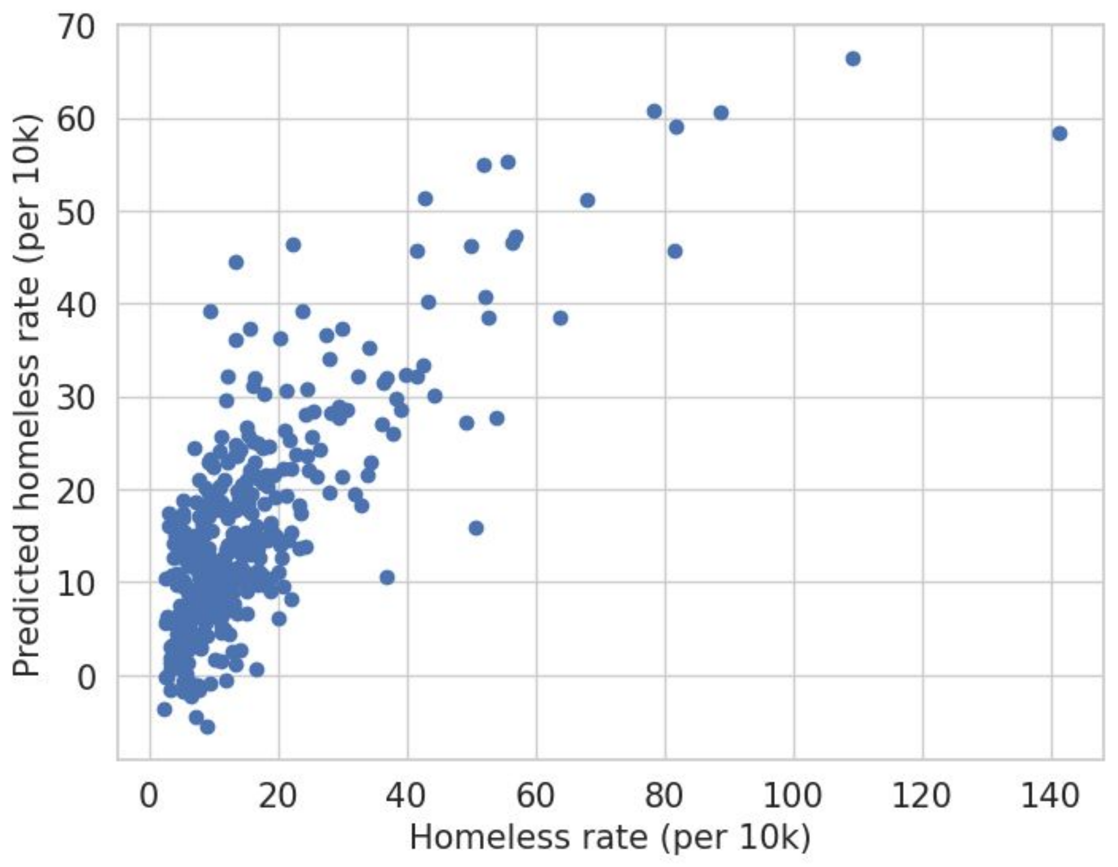
How did I get my data?

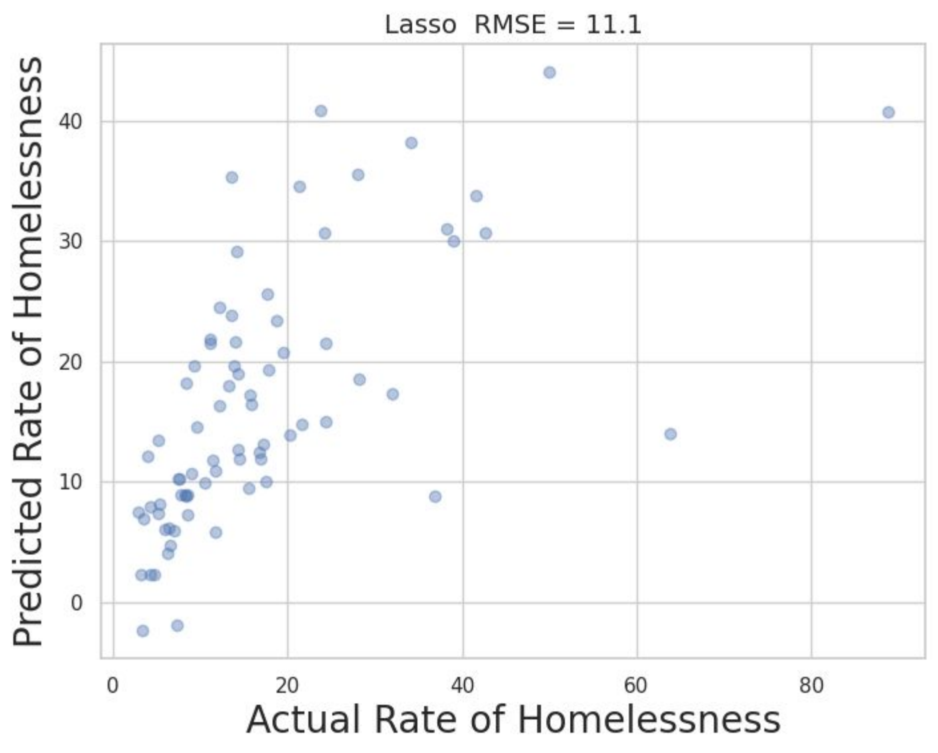
The data for this project are described in HUD’s report Market Predictors of Homelessness

Steps I took:

Firstly, we provided a comprehensive description of the data source, including relevant links to access the data whenever possible. Secondly, we carefully inspected the contents of each data set to gain a thorough understanding of their structure and contents. If needed, we performed data type conversions to ensure the columns were in the appropriate formats. To ensure data cleanliness, we removed any unnecessary parts from the data sets, such as duplicate entries or out-of-range values. From the available columns, we selected a subset that would serve as predictors in our analysis, omitting those that were not required. We also addressed missing values by either imputing them or removing rows with NaN values, depending on the situation. To conform to best practices, we organized the data frame in a tidy or long format. Additionally, if necessary, we renamed columns to adhere to conventions of lowercase, snake_case naming that is easily understandable. Finally, we created derived variables that would provide valuable insights for our analysis. These steps ensured that our data was appropriately prepared and ready for further analysis.

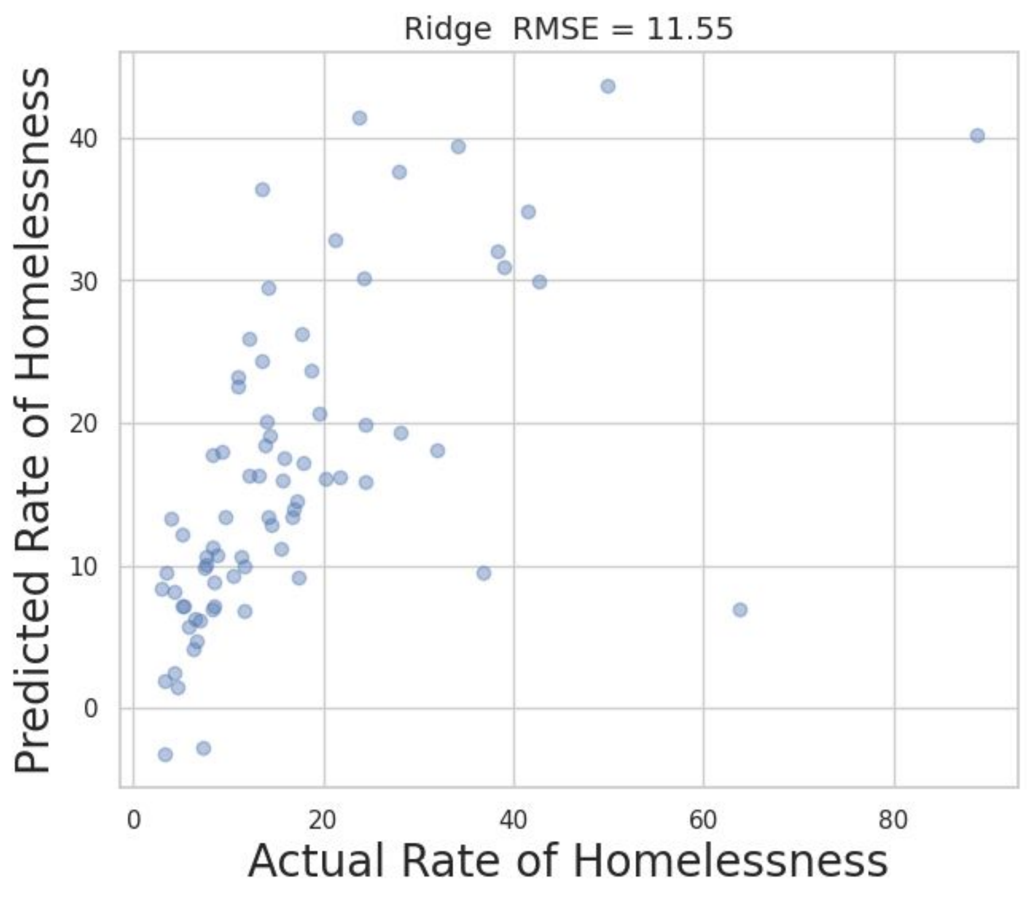
Predictive model designed to estimate Homelessness Rates.



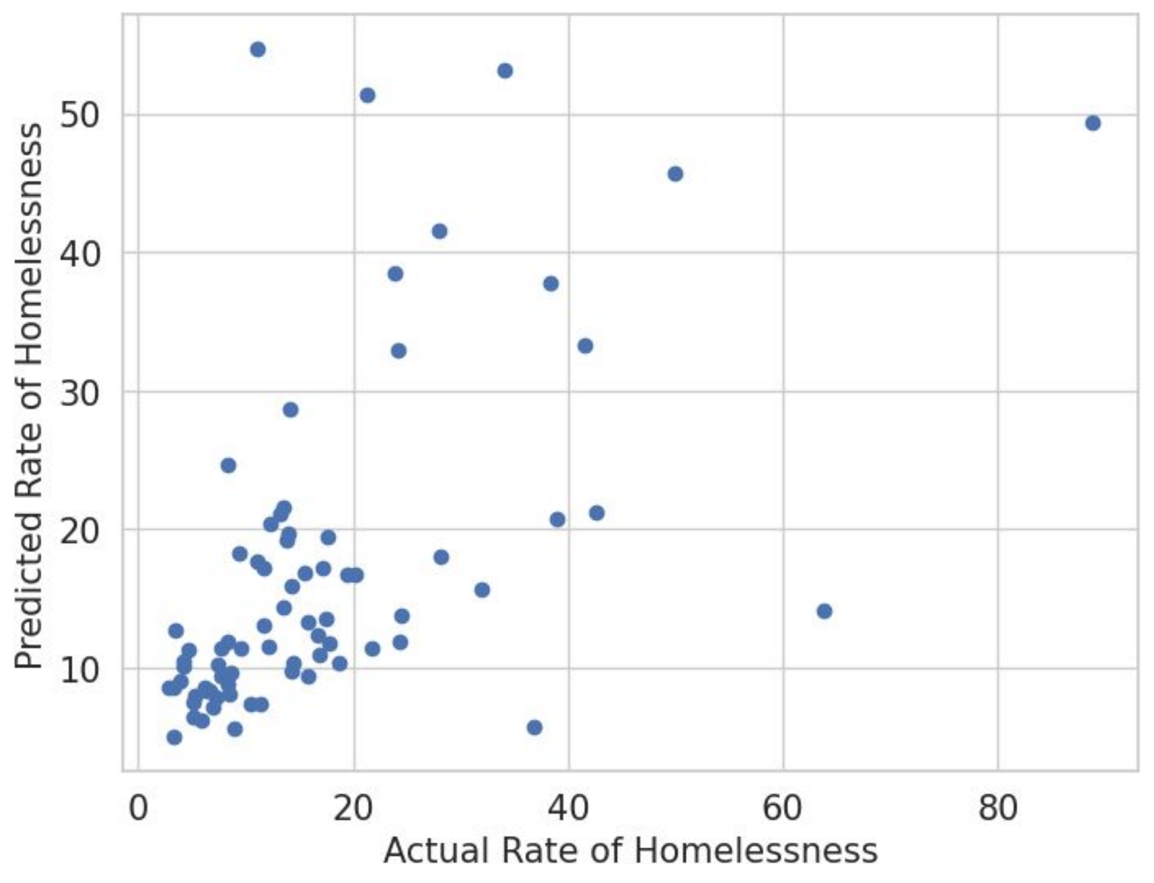


Predictive model designed to estimate Homelessness Rates using Lasso Regression.

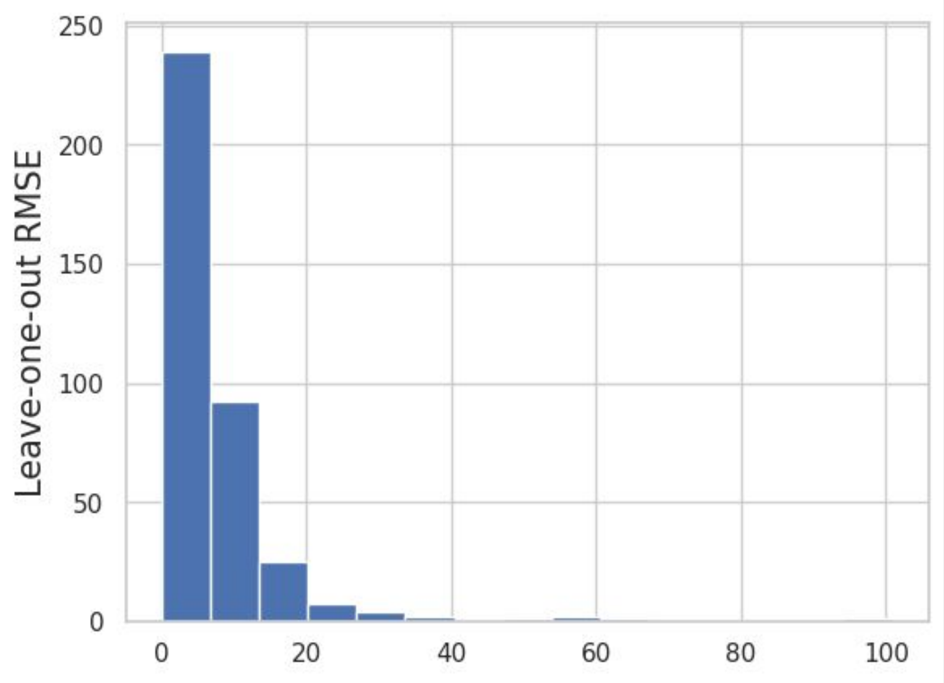
Predictive model designed to estimate Homelessness Rates using Ridge Regression.



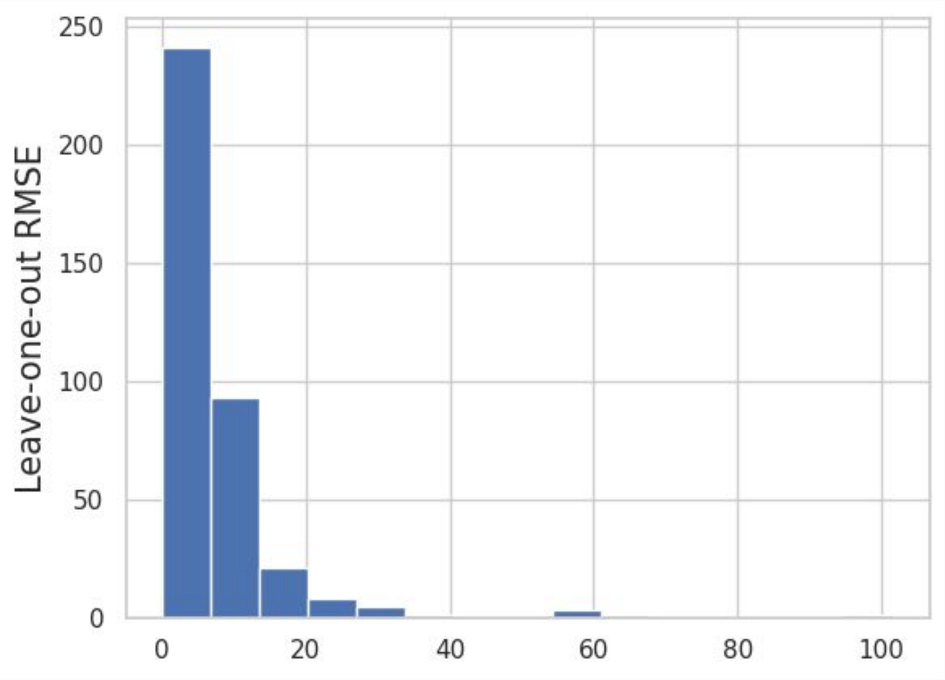
Predictive model designed to estimate Homelessness Rates using XGBoost.



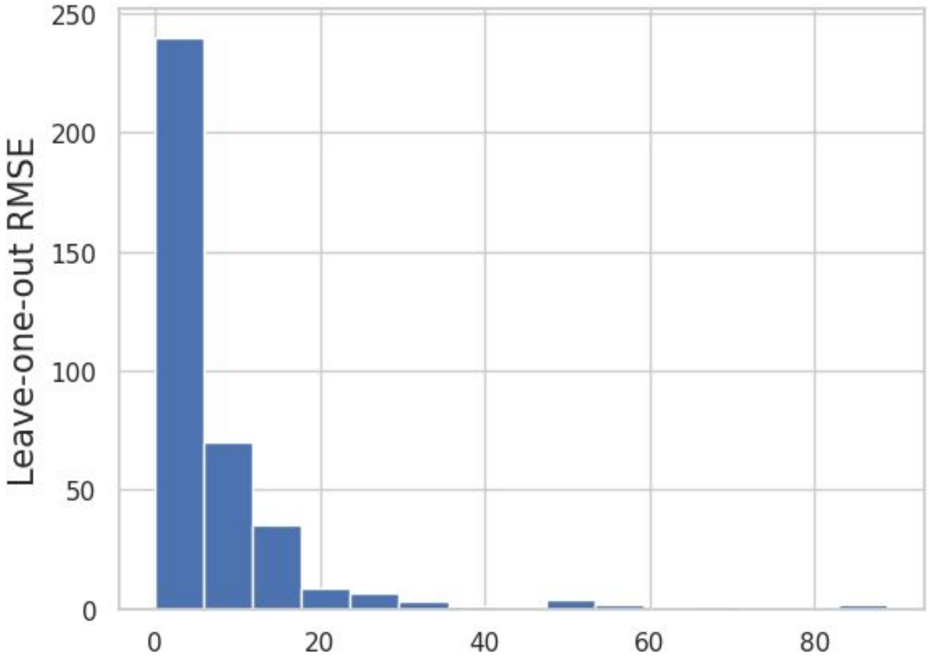
Comparing the RMSE and MAE (Mean Absolute Error) values, we can see that the Lasso regression has the lowest RMSE of 11.097, followed by the Ridge regression with an RMSE of 11.555, and the XGBoost model has the highest RMSE of 12.943. Lower values of RMSE and MAE indicate better model performance, as they indicate smaller prediction errors. Based on these metrics, the Lasso regression model appears to have the best performance among the three models, as it has the lowest RMSE and the highest R-squared.



The coefficients in the Lasso regression model indicate the influence of each feature on the model's predictions. Larger coefficients mean that a feature has a stronger impact on predicting the target variable, while smaller coefficients indicate a weaker impact.



When comparing the Ridge regression model with the Lasso regression model, we find similarities and differences in the coefficients. The intercept term is comparable, indicating a similar starting point. Certain features, like census_region, city_or_urban, and house_price_index_2009, have consistent influence in both models. However, there are notable differences for features such as HUD_unit_occupancy_rate, high_housing_density, and share_overcrowded_units_2016, which have larger coefficients in Ridge regression, suggesting greater importance in the presence of multicollinearity. Overall, Ridge regression has a smoother regularization effect with fewer zero coefficients, assigning significance to more features compared to Lasso regression.



The importance scores obtained from the XGBoost (Extreme Gradient Boosting) model provide insights into the relative contribution of each feature to the model's predictions. Among the features, "census_region" emerges as the most influential, followed by "share_HUD_units," "share_renters_2016," "percent_population_0_19," and "total_Jan_precipitation," which also exhibit significant importance. Conversely, the feature "suburban" appears to have minimal impact on the predictions, as indicated by its low importance score. Notably, all features have non-zero importance scores, implying that each feature contributes to the model's decision-making process to some extent. By considering these importance scores and the corresponding feature names, valuable insights can be gained regarding the features that hold greater importance in the XGBoost model and their contributions to the overall predictive performance.

The XGBoost Regression model has the lowest RMSE of 2.207, followed by the Lasso Regression model with an RMSE of 5.160, and the Ridge Regression model with an RMSE of 5.940.

Lower RMSE values indicate better model performance, as they indicate smaller deviations between the predicted values and the actual values. Therefore, in terms of RMSE, the XGBoost Regression model seems to outperform the other two models.

Conclusion

In conclusion, our project aims to understand homelessness in the United States using data science methodology. We seek to provide valuable insights by exploring alternative modeling approaches and examining the relationship between predictor variables, including geographical region, and homelessness rates. By analyzing coefficients and importance scores from models such as Lasso regression, Ridge regression, and XGBoost, we gain insights into the effectiveness of incorporating regional differences and interactions. Our findings contribute to ongoing efforts to address and prevent homelessness in the U.S., supporting HUD's objectives and providing valuable information for developing accurate predictive models.